André Eliseu
UAG-UCEH/IETEC

António Horta Branco
IETEC

Um laboratório para a aprendizagem de gramática assistida por computador: o projecto Sócrates

1. O PROJECTO

1.1. OBJECTIVOS

O Projecto "Sócrates" teve como objectivo produzir um protótipo de programa-ferramenta na área do Ensino da Gramática.

Este programa foi concebido como um laboratório de gramática cujas funções básicas são:

- Permitir ao utilizador representar conhecimentos gramaticais, partindo da sua competência linguística e dos conceitos adquiridos na aprendizagem escolar. Ao trabalhar com o programa, o utilizador é levado a aplicar:

  noções de carácter formal, como

  - regra de reescrita
  - entrada lexical

- recursividade
- relação de dominação
- relação de precedência

noções de carácter descritivo, como:

- categoria sintáctica (categorias nucleares e sintagmáticas)
- selecção sintáctica
- selecção semântica
- função sintáctica
- concordância
- flexão (nominal e verbal)
- estrutura de constituintes

Ao trabalhar com o programa, o utilizador é, igualmente, levado a aplicar uma terminologia linguística padronizada.

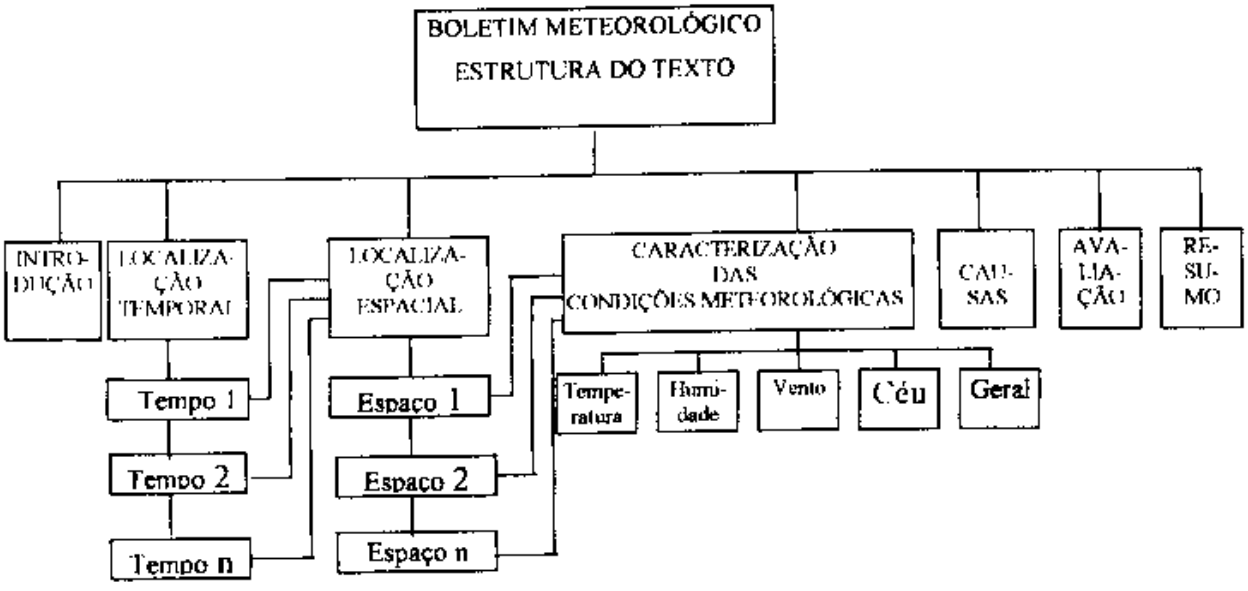
- Permitir ao utilizador explorar a informação linguística que tiver sido introduzida no sistema, testando as gramáticas e os dicionários disponíveis através da aplicação a diferentes expressões linguísticas. Pelo processo de construção e/ou experimentação, o utilizador será levado a desenvolver estratégias de optimização da representação formal dos conhecimentos linguísticos e, desta forma, a contactar com as propriedades formais de uma gramática.

- O utilizador será levado também a adquirir, consolidar ou desenvolver conceitos linguísticos de carácter descritivo e a aperceber-se que a gramática de uma língua é uma teoria que estabelece predições testáveis sobre as propriedades dessa língua.

superestrutura para este tipo textual, seguindo o exemplo de van Dijk, que, em *News as Discourse* (van Dijk, 1988), levou mais longe uma primeira indicação incluída no seu livro fundamental sobre as macrostruturas (van Dijk, 1980), ao apresentar uma hipótese de superestrutura para a notícia na imprensa escrita. Bell (1991) apresentou um modelo algo mais pormenorizado, que também foi tido em conta.

Debruçando-nos sobre a estrutura do texto, verificámos que os textos estudados apresentam uma mesma superestrutura subjacente (cf. Figura 1). Todos os textos seguem este esquema mas podem omitir algumas categorias bem como variar a ordem da sua apresentação. O esquema permite detectar as categorias contempladas no texto, a distribuição da informação pelas mesmas e as que não são preenchidas (cf. Apêndice).

Figura 1



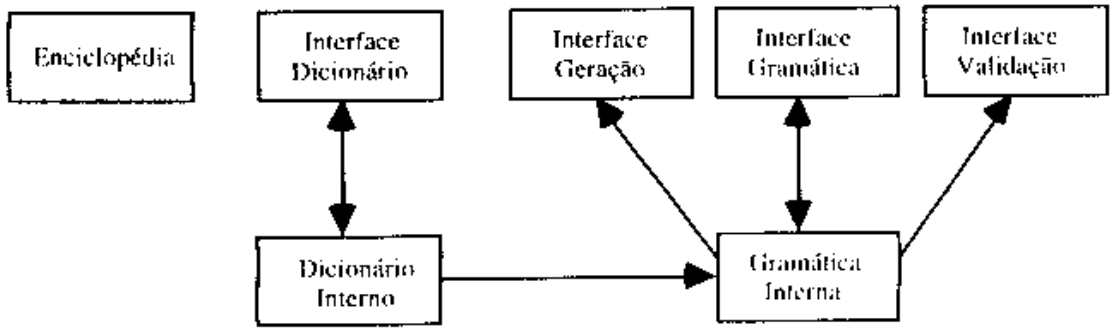
A categoria INTRODUÇÃO, por ex., foi quase sempre omitida na SIC e na RTP. Apenas na TVI encontrámos um primeiro espaço introdutório que inclui uma chamada de atenção para as primeiras imagens (ver Apêndice). O texto propriamente dito resulta da apresentação cronológica dos vários estados do tempo, identificando o local

protótipo sido desenvolvido sobre o AV-Parser, na versão de 1992, da autoria do Prof. Mark Johnson da Brown University.

O AV-Parser é um ambiente de desenvolvimento dedicado para aplicações na área da Linguística Computacional que disponibiliza um formalismo da família dos formalismos de unificação. Uma parte importante do trabalho desenvolvimento informático consistiu portanto na construção de um compilador que reescreve no referido formalismo as especificações linguísticas constantes da gramática e do dicionário em utilização. Estes últimos são construídos através da utilização dos instrumentos formais de mais alto nível, adequados aos utilizadores-alvo da aplicação, que são apresentados na secção seguinte.

2.2. FUNCIONALIDADES

A estrutura geral do protótipo pode ser representada pelo seguinte diagrama (representando os módulos principais e o fluxo de informação entre eles).



O sistema funciona interactiva e modularmente, sendo possível testar as representações (sintácticas e lexicais) introduzidas; a cobertura lexical e sintáctica será alargada à medida das necessidades e interesses do utilizador, dentro dos limites da cobertura linguística definidos pelo protótipo (cf. 2.2).

O interface de utilizador permite a caracterização de items lexicais, a identificação das categorias, a definição de estruturas de constituintes e a expressão das relações sintácticas através da manipulação dos elementos do écran (menus, botões e paletas de ferramentas).

Como se disse, o programa pode funcionar em situação de auto-aprendizagem ou de aprendizagem apoiada, isto é, o utilizador pode ir construindo o dicionário e a gramática, aumentando a cobertura e a complexidade do sistema de acordo com a sua progressão de aprendizagem e com os seus objectivos ou pode trabalhar com gramáticas e um dicionários previamente definidos e usar o sistema para obter soluções para problemas pré-definidos ou verificar as suas próprias soluções.

2.2.1. "DICIONÁRIO"

O módulo "Dicionário" permite criar entradas lexicais através do preenchimento de fichas contendo uma forma (correspondente à palavra que se quer introduzir) e uma categoria sintáctica (nome, verbo, preposição, etc). Para cada categoria seleccionada, a ficha exibe os traços morfo-sintácticos relevantes (por exemplo, "Número", "Género", etc. para Nome, "Modo", "Tempo", etc. para Verbo) e, se for caso disso, as propriedades de selecção da forma introduzida.

Uma vez que o programa não dispõe de um analisador morfológico, é necessário introduzir cada forma distinta (quanto a Número, Tempo, etc.) que se pretenda ter no "Dicionário".

Note-se que o programa não é fornecido com qualquer dicionário previamente carregado, cabendo ao utilizador definir o seu léxico. No caso de o programa ser utilizado na modalidade de "aprendizagem apoiada", o utilizador poderá, evidentemente, trabalhar com um dicionário construído por outrem, possivelmente, o professor.

2.2.2. "GRAMÁTICA"

O módulo "Gramática" permite criar representações sintácticas sob a forma de diagramas em árvore, cujos nós são seleccionados a partir de uma lista fechada fornecida pelo programa. Cada uma destas árvores constitui a expressão de uma regra da gramática.

Note-se que o programa não é fornecido com qualquer gramática previamente definida, cabendo ao utilizador construir a sua gramática, excepto no caso de o programa ser utilizado na modalidade de "aprendizagem apoiada", em que o utilizador trabalhará com uma gramática já construída.

Para efeitos de exposição, vamos considerar que a tarefa do utilizador é definir uma gramática e alargar a sua cobertura. Tal poderá ser feito através de uma de duas estratégias:

A — a estratégia mais simples consiste em acumular tantas representações completas quantas as diferentes estruturas (frases ou sintagmas) que se quer tratar; por exemplo, a regra 1 para frases com um verbo transitivo, a 2 para frases com um verbo intransitivo, etc.

Vantagens: acessibilidade a qualquer utilizador com os conhecimentos mínimos pressupostos.

Desvantagens: elevado grau de redundância na gramática.

B — em alternativa, pode optar-se pela construção de regras sem definir a estrutura interna de todos os nós representados, utilizando a capacidade do programa para expandir esses nós (neste caso, o programa inserirá na representação a subestrutura relevante, desde que existam na "Gramática" as regras que descrevem tais nós). Por exemplo, regras que definem os constituintes imediatos de Frase, regras que descrevem a constituição interna dos SNs, etc.

Vantagens: maior elegância da gramática, maior economia de escrita da gramática, maior controlo da interacção das regras.

Desvantagens: requer um nível de conhecimentos acima do mínimo.

2.2.3. "GERAÇÃO"

O módulo "Geração" produz a descrição da estrutura de constituintes de uma sequência de palavras introduzida pelo utilizador, na condição de as palavras constarem — com a categoria relevante – do dicionário em utilização e de existirem na "Gramática" em uso as regras capazes de descrever essa expressão.

2.2.4. "VALIDAÇÃO"

O módulo "Validação" verifica a conformidade das descrições estruturais (associadas a expressões) introduzidas pelo utilizador relativamente a uma gramática e a um dicionário em utilização.

2.2.5. "ENCICLOPÉDIA"

A "Enciclopédia" é um módulo de auxílio ao utilizador que fornece definições de termos linguísticos. A informação é acessível a partir de um índice (lista de entradas) ou de palavras-chave no interior do corpo das definições.

2.3. COBERTURA LINGUÍSTICA

Os módulos "Gramática" e "Dicionário" cobrem as noções de uma gramática elementar do Português, nos domínios da Morfo-sintaxe e da Sintaxe, abrangendo os conteúdos dos tópicos dos programas escolares nestes campos. O módulo "Enciclopédia" contém informação sobre tópicos linguísticos que o utilizador pode consultar se dela necessitar para trabalhar com o programa, servindo como um manual de gramática electrónico.

2.3.1. CATEGORIAS SINTÁCTICAS

As categorias sintácticas incluídas na "Gramática" são as seguintes:

Categoria nuclear	**Categoria sintagmática**
----	F frase

A	adjectivo	SA	sintagma adjectival
ADV	advérbio	SADV	sintagma adverbial
ART	artigo	SDET	sintagma determinante
AUX	auxiliar	SV	sintagma verbal
CONJ	conjunção	não define um sintagma, é sempre dominada por F	
DEM	demonstrativo	SDET	sintagma determinante
N	nome	SN	sintagma nominal
NUM	numeral	SDET	sintagma determinante
P	preposição	SP	sintagma preposicional
PESS	pronome pessoal	SN	sintagma nominal
POSS	possessivo	SDET	sintagma determinante
QUANT	quantificador	SDET	sintagma determinante
V	verbo	SV	sintagma verbal

2.3.2. FUNÇÕES SINTÁCTICAS

O utilizador pode igualmente definir funções sintácticas atribuindo etiquetas aos nós de uma árvore. As funções sintácticas admitidas pelo programa são as seguintes:

> sujeito, objecto directo, objecto indirecto, oblíquo, nome predicativo do sujeito, nome predicativo do objecto directo.

2.3.3. RELAÇÕES CONFIGURACIONAIS

Na "Gramática" são definidas relações de dominação e precedência entre nós. A relação de dominação é expressamente atribuída pelo utilizador através da operação de construção da árvore; a relação de precedência é inferida a partir da posição relativa dos nós na árvore.

2.3.4. CONCORDÂNCIA

É possível definir relações de concordância entre nós de uma árvore, através de um processo em que o utilizador assinala os nós envolvidos e o sistema atribui automaticamente índices a esses nós. A verificação da concordância faz-se no módulo de "Validação": as palavras que são dominadas por nós com um mesmo índice de concordância têm que possuir na sua entrada no "Dicionário" os mesmos valores para os traços de número (ou género e número, consoante as respectivas categorias).

2.3.5. TIPOS DE REGRAS E CONSTRUÇÕES

A "Gramática" permite descrever estruturas sintácticas simples e estruturas sintácticas complexas com as limitações abaixo indicadas, num único nível de representação.

Apenas podem ser formuladas regras de tipo sintagmático. Não é possível definir regras de tipo transformacional, isto é, qualquer tipo de regras que derivem uma representação a partir de outra representação.

A "Gramática" é capaz de tratar árvores contendo nós não expandidos, inserindo a sub-estrutura relevante na representação e de lidar com a recursão.

Não é possível validar ou gerar construções com categorias vazias, isto é, em que exista um nó terminal que não domine nenhum elemento lexical.

2.3.6. SUBCATEGORIZAÇÃO E RESTRIÇÕES DE SELECÇÃO

As restrições de selecção e as propriedades de subcategorização dos verbos podem ser declaradas na entrada lexical de cada item. Esta informação pode ser usada para impor condições sobre a aplicação de regras sintácticas definidas na gramática.

2.3.7. FILTROS

Embora mantendo o princípio de que cabe ao utilizador definir o conteúdo das regras que escreve, considerou-se útil e instrutivo bloquear a possibilidade de infringir certos princípios gerais sobre os quais assenta o tipo de gramática que o programa assume. Por isso, o módulo "Gramática" contém internamente um conjunto de restrições que incidem exclusivamente sobre aspectos formais da construção das regras sintácticas, como por exemplo, a obrigatoriedade de um nó terminal ser sempre dominado por um outro nó, a

impossibilidade de um dado nó poder ser um nó filho de mais de um nó ou as restrições sobre os constituintes imediatos de dadas categorias sintagmáticas.

Note-se que estes filtros não garantem a adequação das regras escritas pelo utilizador. De acordo com a estratégia pedagógica do "Sócrates", será através do funcionamento dos módulos de "Geração" e "Validação" que o utilizador, por comparação com as suas expectativas, verificará tal adequação.

3. A APRENDIZAGEM DA GRAMÁTICA ASSISTIDA POR COMPUTADOR

3.1. O SÓCRATES NO CONTEXTO DO EALLAC[3]

Tomando em consideração o domínio das aplicações para o ensino assistido por computador, o Sócrates apresenta um conjunto de características específicas e inovadoras que interessa colocar em destaque.

Em primeiro lugar notamos que, de forma diferente do que acontece com a esmagadora maioria das aplicações para a área da linguagem, o Sócrates não tem por âmbito o ensino de uma língua, mas sim o ensino da Linguística.

Segundo, o Sócrates não se limita a fazer uma mera transposição para o ambiente computacional de metodologias de ensino que têm a sua expressão original em suportes

---

3 Para uma apresentação sinóptica do EALLAC, vd. Branco & Mendes, "Ensino de Línguas e Linguística Assistido por Computador", in Mateus & Branco (orgs.), Engenharia Linguística - Actas do Curso "Engenharia Linguística" da Universidade de Verão "Estudos Gerais da Arrábida - Conferências do Convento", Lisboa, Colibri, no prelo.

tradicionais, como é o caso, por exemplo, das inúmeras aplicações com exercícios de escolha múltipla, exercícios de preenchimento de lacunas, etc. O Sócrates é um utilitário que apresenta potencialidades didácticas que só é possível obter por via da exploração das funcionalidades específicas do computador e dos resultados da investigação básica em Linguística Computacional.

Terceiro, o Sócrates afasta-se da tendência geral de encarar os utilitários didácticos segundo a perspectiva do "plug and play", a qual leva a que as aplicações se encontrem limitadas pela tentativa de as mesmas se assemelharem o mais possível aos jogos de vídeo de diversão, com pontuação, bónus, tabela dos melhores jogadores, etc. Com efeito, o Sócrates não é de todo um jogo, nem mesmo um ambiente para a aprendizagem auto-suficiente do utilizador. Para tirar partido das potencialidades didácticas únicas desta aplicação é preciso entender que a mesma está provavelmente para o ensino da Linguística como o laboratório de química está para o ensino da Química, constituindo portanto um instrumento a utilizar por docentes e alunos no processo de ensino-aprendizagem.

O conjunto destas características leva a que se classifique o Sócrates, no quadro do domínio do Ensino/Aprendizagem de Línguas e Linguística Assistido por Computador (EALLAC), como uma aplicação dedicada de segunda geração, baseada em técnicas de Engenharia Linguística.

3.2. SÓCRATES: MODOS DE UTILIZAÇÃO DIDÁCTICA

O programa pode ser utilizado em diferentes contextos de aprendizagem, que designaremos por "modalidades de utilização": **auto-aprendizagem** e **aprendizagem apoiada**.

- na primeira, o utilizador-aluno constrói uma "Gramática" e um "Dicionário", de acordo com um objecto de aprendizagem que ele próprio define, podendo alargar a sua cobertura e alterar o seu conteúdo consoante os resultados dos testes a que os

submete. O aluno tem acesso aos módulos "Gramática", "Dicionário", "Validação" e "Geração".

- na segunda modalidade, o utilizador-aluno trabalha com uma "Gramática" e um "Dicionário" construídos por outrem (o professor) com o objectivo de dar conta de um conjunto restrito de construções. Neste caso, o aluno apenas acede aos módulos "Validação" e "Geração" tendo como objectivo verificar as soluções de problemas apresentados.

O programa "Sócrates" foi concebido para servir como um laboratório que permite ao utilizador explorar a informação linguística fornecida ao sistema. Como tal, impõe um número mínimo de restrições sobre o conteúdo das regras sintácticas (designadas "filtros") ou das especificações lexicais (incorporadas no formato das fichas do Dicionário). O utilizador pode não só testar a gramática e o dicionário que construiu , como explorar as propriedades do sistema criado.

As características acima enunciadas decorrem dos dois princípios básicos em que assenta a concepção do programa:

- O programa é um laboratório de gramática. A flexibilidade na definição do conteúdo linguístico dos módulos da "Gramática" e do "Dicionário", impondo um número mínimo de restrições, teve em vista introduzir experimentalmente a reflexão sobre algumas características básicas das gramática das Línguas Naturais.

- O programa pretende levar o utilizador a *aprender* um certo número de conceitos, técnicas e nomenclatura linguísticos *a partir da tarefa de construção* da "Gramática" e do "Dicionário".

Para ilustrar as modalidades de utilização do programa vamos considerar que ele pode ser aplicado a dois tipos de objectos:

(1) a. Língua Natural, presumivelmente o Português

b. "Pseudo-língua", criada pelo utilizador.

e usado em dois domínios de exploração de conhecimentos

(2) a. Adequação descritiva do léxico e da gramática construídos.

b. Observação de propriedades do léxico e de uma gramática de constituintes

Naquele que poderemos presumir como o uso básico do programa, o utilizador toma como objecto o Português (1a), construindo um "dicionário" de formas (palavras existentes no léxico da língua e respectivas propriedades categoriais e de selecção) e uma "gramática" que dê conta de um certo número de construções pertencentes à sintaxe do Português, envolvendo combinações daquelas formas lexicais. A adequação empírica do dicionário e da gramática (2a) será verificada através da sua aplicação a diferentes conjuntos de dados. Esta modalidade de utilização visa sobretudo a aprendizagem de propriedades lexicais e sintácticas do Português, o domínio de técnicas de análise sintáctica e da nomenclatura gramatical.

No entanto, a versatilidade do programa permite a sua utilização tendo em vista outros objectivos. Assim, pode explorar-se a possibilidade de tomar como objecto de descrição outra Língua Natural — possibilidade que faz todo o sentido desde que as categorias lexicais e sintácticas do programa correspondam às do fragmento de gramática que se quer descrever. Outros exemplos serão construir um sistema de regras diferente das do Português e aplicá-lo ao "Dicionário" português ou usar a "Gramática" do Português e alterar a classificação no "Dicionário", observando o resultado (via módulo de "Geração") e confrontando-o com as intuições do utilizador. Pode igualmente trabalhar-se com uma "língua" artificial (1b), procurando estabelecer e descrever um conjunto de regras e definições de dicionário.

Numa outra perspectiva (2b), pode utilizar-se o programa para explorar certas propriedades lexicais e sintácticas. Por exemplo, pode partir-se de uma "Gramática" que

reflicta os conhecimentos escolares do utilizador e observar os efeitos da eliminação de regras (quais as regras necessárias para descrever um dado conjunto de frases? que tipo de construções acrescentadas ao conjunto inicial requer a introdução de novas regras? etc.). Ou, explorando a caracterização lexical: que propriedades podem ser omitidas preservando a adequação das representações sintácticas?

Finalmente, a "Enciclopédia" é um manual de gramática do Português, organizado em fichas que contêm as noções gramaticais cujo conhecimento é necessário para compreender e trabalhar com o programa e pode ser usada independentemente dos restantes módulos, como uma base de dados linguísticos utilizável em qualquer contexto de aprendizagem[4].

André Eliseu
UA-UCEH/ILTEC
email: ase@iltec.iltec.pt

António Horta Branco
ILTEC
email: ahb@iltec.iltec.pt

---

[4] A cobertura e o tratamento dos temas na "Enciclopédia" excedem o mínimo requerido pela sua utilização como auxiliar dos programas escolares, fornecendo informação básica sobre tópicos da gramática do Português, sobre nomenclatura gramatical e sobre propriedades formais das gramáticas das línguas naturais. Os temas que se considerou serem mais problemáticos ou estarem insuficientemente tratados na bibliografia escolar foram objecto de um tratamento mais extenso.